知識発見のためのテキストマイニング技術

4N－6

那須川 哲哉　長野 徹
日本アイ・ビー・エム株式会社 東京基礎研究所

1 はじめに

日々膨大な量の情報が文書形式で記述蓄積されている。文書形式の記述は自由度が高いため、例えば顧客の情報を記録しようという場合、選択肢を指定する形式で記述するよりも、はるかに詳細で豊富な情報を記録することが可能になる。その反面、文書形式のデータは分析が困難であり、基本的には人が目を通して分析するしか手段がないのが現状である。分析の労力を軽減させるための技術としては、分析対象を絞るための検索技術や、大量の文書を分類整理するためのクラスタリングやクラシフィケーションの技術が存在する。しかし、これらの技術のみでは、原文に全く目を通さずに何らかの問題を発見するというような知識発見の自動化にはつながらない。

本稿では、膨大な文書データの中から注目に値すべき内容を自動的に発見する手法を提案し、実データへの適用結果を通してその有効性を示す。

2 文書データからの知識発見

2.1 相関ルールの導出

大規模データからの知識発見としては、データマイングにおける相関ルールの導出[1]が良く知られている。文書データに関しても同様の技術を適用し、言葉と言葉の相関ルールを導出する事で何らかの知識が得られるのではないかとう期待が生じる。ところが実際には、文書データではアイテムとなる言葉の種類が膨大な数に及ぶため、導出される相関ルールの数も膨大になり、その上、雑多な内容が混合されているため、この中から有用なルールを見出すのは非常に困難である。

実際の実験結果を見てみると、例えば「お茶の水」が一語として辞書登録されていない場合に、「茶」と「水」の相関ルールが導出されるというように、知識発見という言葉からイメージされるような結果はなかなか得られない。

TAKMI: Text Analysis and Knowledge MIning
Tetsuya Nasukawa, and Tohru Nagano
IBM Research, Tokyo Research Laboratory

2.2 観点の導入

従って、文書データから有効な知識を抽出するためには、雑多で膨大な言葉を全て混合して分析するのではなく、どのような知識を得たいかにより、対象を限定する必要がある。

文書データは一般的に多様な解釈が可能であり、利用者の観点によって全く異なる知識を引き出すことができるという特徴がある。そのため、文書データからの知識発見にあたっては、利用者の観点の取り込みが必要となってくる。

例えば、自社製品に関する顧客からの問合せデータから知識を抽出する場合、製品の信頼性に関する責任者であれば『製品の不具合がないかどうか』に、新製品の開発に関する責任者であれば『顧客が特に何を喜ぶか』に関心を持つと考えられる。そこで、例えば、『製品の不具合がないかどうか』を調べたい場合、文書中から、製品に関する概念（製品名など）と、不具合（問題）に関する概念を抽出し、その相関関係を調べることで、どの製品にどのような問題が多いかを調べることができる。

2.3 特異性の検出と分析

しかし、全ての製品に関して、逐一分析を行うのは大変な作業である。その上、例えば、ある製品に対して「操作が難しい」というコメントが存在したとしても、全ての製品に関して同じようなコメントが存在するなら、特にその製品に注目する必要はないという判断も可能である。

従って、注目すべき内容をいかにして自動的に認識すべきかが重要であり、そのための判断基準が必要となる。ここで、一般的に類似概念（例えば同じ種類の製品）に関しては同じような性質が現れると仮定する。例えば、どのような製品に対しても程度の差こそあれ「使いにくい」「動かなくなる」「傷つく」といった問合わせが同じように現れる傾向があると仮定する。この仮定を前提とすれば、特定の

	スペック一教える	メモリー一増設する	ISDN一接続する	スペック一知る	アップグレードサービス一受ける	国際保証一教える
APTIVA	29 (0.44%)	46 (0.7%)	8 (0.12%)	7 (0.11%)	4 (0.06%)	0 (0.0%)
THINKPAD	32 (1.37%)	9 (0.38%)	0 (0.0%)	7 (0.3%)	0 (0.0%)	1 (0.04%)
PS/V	3 (0.54%)	12 (2.15%)	0 (0.0%)	4 (0.72%)	0 (0.0%)	0 (0.0%)

図1： 各ブランドの機器に関するお客様の要望の特徴の表示結果

内容の出現傾向が極端な個所に注目すれば、そこに意味のある情報が含まれている可能性が高いと判断する事ができる。すなわち、ある概念とある概念との相関関係の特異性を調べることで、注目すべき内容の自動的な検出が可能になる。

この処理を適用するには、文書中の各表現が基本的にどのような性質の概念を示しているかを考慮して文書中から情報抽出を行う必要がある。したがって、単なる文字列としてのキーワードを抽出するよりも深い言語処理が必要である。[2]

文書データには豊富な情報が含まれているため、注目すべき内容を検出した後には、その内容を含む文書を対象とした分析を行う事により、さらに詳細な知識を得る事が可能になる。例えば、対象文書中の出現概念の分布傾向を全文書データ中の分布傾向と比較する事で、検出された内容に関連性の強い概念を抽出することができる。

2.4 実現例

この考え方に基づく特異性検出機能を、テキストマイニングのプロトタイプシステム TAKMI[2]上に実現した。TAKMI では文書中に記述された概念をカテゴリ分けして抽出しており、同じカテゴリに属する概念は同じような性質を取るものとして扱う事ができる。ユーザが二つのカテゴリ（仮にＡとＢとする。）を指定すると、カテゴリＡに属する各概念に対し、その概念と共起しているカテゴリＢの各概念の割合を求め、その割合が他と異なるものを出力する。またインタラクティブな分析を行うＧＵＩ上では、図１のように二つのカテゴリをそれぞれ縦軸と横軸にとり、マトリクス状に表示した上で、割合の特異な部分をハイライトする。さらに、こうして着目すべき内容を表示した後、ハイライトされた部分を指定する事により、その内容を含む文書集合に対象を絞り込み、異なる観点からの分析を行うことで、より詳細な知識を獲得する事ができる。

3. 適用事例

本手法を実際のデータに適用し、その有効性を調べた。対象データは日本ＩＢＭ（株）お客様相談センターに電話で寄せられた問合せの内容を文書形式で記録した報告書である。

図１には、98年１月から５月までの期間内で総合案内に分類されたデータのうち、Aptiva、ThinkPad、PS/V という３つのブランドの機器に関する問合せ約１万件（9356 件）の中に含まれているブランド別の要望の特徴が示されている。各セルにおいて左側の数値が問合せの件数を示し、括弧内の数値は左端に表示されているブランドに関する問合せの中での比率を示している。問合せ総数の比は、Aptiva : ThinkPad : PS/V がおよそ 10 : 3 : 1 となっている。そのため、例えば「メモリを増設したい」という要望を見ると、絶対数では Aptiva に関する問合せが 46 件と一番多いが、構成比では PS/V に関する問合せにおける比率が 2.15%と他より特に高い値になっている。したがって、PS/V に関する「メモリを増設したい」という問合せのセルがハイライトされている。

同様にして、月単位のデータに本手法を適用し、機種別の問題の分布を分析してみると、いくつかの機種で特定の問題がハイライトされる結果が得られた。例えば、ある機種に対しては「遅い」という問題を指摘される割合が他の機種に比べて特に高く、その問合せに関連性の高いハードウェアはハードディスクであるという結果が得られた。実際にこの機種においては一部のハードディスクにおいてシールディングの問題が発見されており、本手法が製品の問題の発見に役立つ事が検証されている。

参考文献

[1] Agrawal,R., Imielinski,T., and Swami,A.: "Mining Association Rules between Sets of Items in Large Databases," In Proceedings of the ACM SIGMOD '93, pp.207-216 (1993)

[2] 那須川，諸橋，長野「テキストマイニング―膨大な文書データの自動分析における知識発見―」情報処理，pp.358-364 (1999)